Miles per gallon Analysis of Motor Trend Data

Jagdeep S. Sihota

Executive Summary

The objective of the study was to look at a data set of a collection of cars and exploring the relationship between a set of variables and miles per gallon (MPG). The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). Data contains 32 observations on 11 variables (See Table 1.1 in the Appendix for more details), and can be found at http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html We were particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

While initial analysis shows that manual transmissions were better for MPG, but regression analysis showed MPG mostly depends on weight and horsepower and transmission was not major factor in determining MPG. Based on this analysis, we can't quantify the MPG difference between automatic and manual transmissions becasue more useful variables are weight, and horsepower.

Exploratory Analysis

To begin to understand the relationship between the variables in the data set and MPG a scatter plot matrix was created to compare all of the variables (Figure 1.1). To explore the relationship between MPG and transmission type, a box plot was created (Figure 1.2). The box plot showed Automobiles with a manual transmission had a higher MPG than those with an automatic.

But there are 10 predictor variables in the data set, so an analysis of variance model is performed, see Table 1.2. The results showed that MPG was most highly correlated with weight (-0.87), the number of cylinders (-0.85), displacement (-0.85), and horsepower (-0.78).

Statistical Modeling

To address the main research question of the relationship between MPG and transmission type, we used simple least squares regression, with MPG as the outcome variable (Y) and transmission as the predictor (X). Table 2.1 shows the results of this regression analysis. The coefficient estimate for manual transmission is 7.24, which tells us that automobiles with a manual transmission get about 7.24 more gallons per mile than those with an automatic transmission. This finding is statistically significant as $p < 0.05$ and the 95% confidence intervals for the estimate were (3.64, 10.85).

The multiple R-squared value is only 0.35, which indicates that only 35% of the variance in MPG is explained by the type of transmission, leaving 65% of the variance unexplained. We looked at additional variables: weight, and number of cylinders. We can see that these are also very influential, and should be included in the model Table 2.2.

Finally, we created a Model with two statistically significant predictor variables, weight and horsepower Table 2.3. The coefficient for weight is -3.88, meaning that on average and with horsepower being held constant, a one unit increase in weight (1000 lbs) equates to a -3.88 lower MPG. This coefficient is statistically significant with a $p<0.05$ and the 95% confidence intervals are (-5.17, -2.58). The coefficient for horsepower is also negative at -0.03, which indicates that on average and holding weight constant, a one unit increase in gross horsepower will result in a .03 loss in MPG. This estimate is statistically significant ($p<0.05$) with 95% confidence intervals of (-0.05, -0.01). The R-squared value of this model is 0.83, which means that 83% of the variance in MPG is explained by the weight and horsepower of the automobile.

Diagnostic Plots

To verify the conditions of validity of our final model we can look at some diagnostic plots of the residuals. The plot of Residuals vs. Fitted values is slightly U-shaped which shows that there may be some non-linearity in the relationship between MPG and one of the variables in the model, suggesting that adding a quadratic term may improve the fit of the data, see Figure 3.1.

After adding quadratic terms for weight and horsepower, we see that the new coefficients are both statistically significant ($p < 0.05$) and the R-squared value (0.89) was the highest among all models, see Table 8. Furthermore, if we look at the diagnostic plot for Residuals vs. Fitted values for this new model, we see that there is no longer a discernible pattern in the variance of the residuals in relation to the fitted values, see Figure 3.2.

Conclusions

By looking at several possible models, we see a relationship does exist between fuel efficiency and transmission type, but that could also be explained by other factors such as vehicle weight and horsepower.

Appendix

Table 1.1

"Motor Trend Data, 1974"

name	description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Figure 1.1: Scatter Plot Matrix

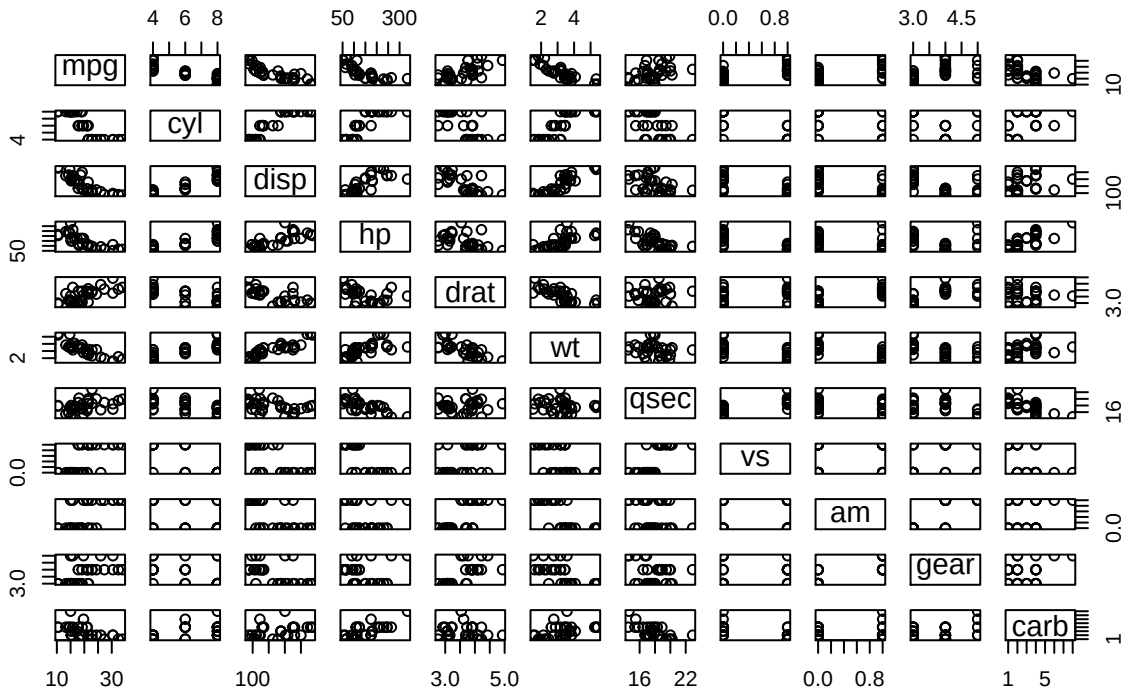

Figure 1.2: MPG by Transmission Type

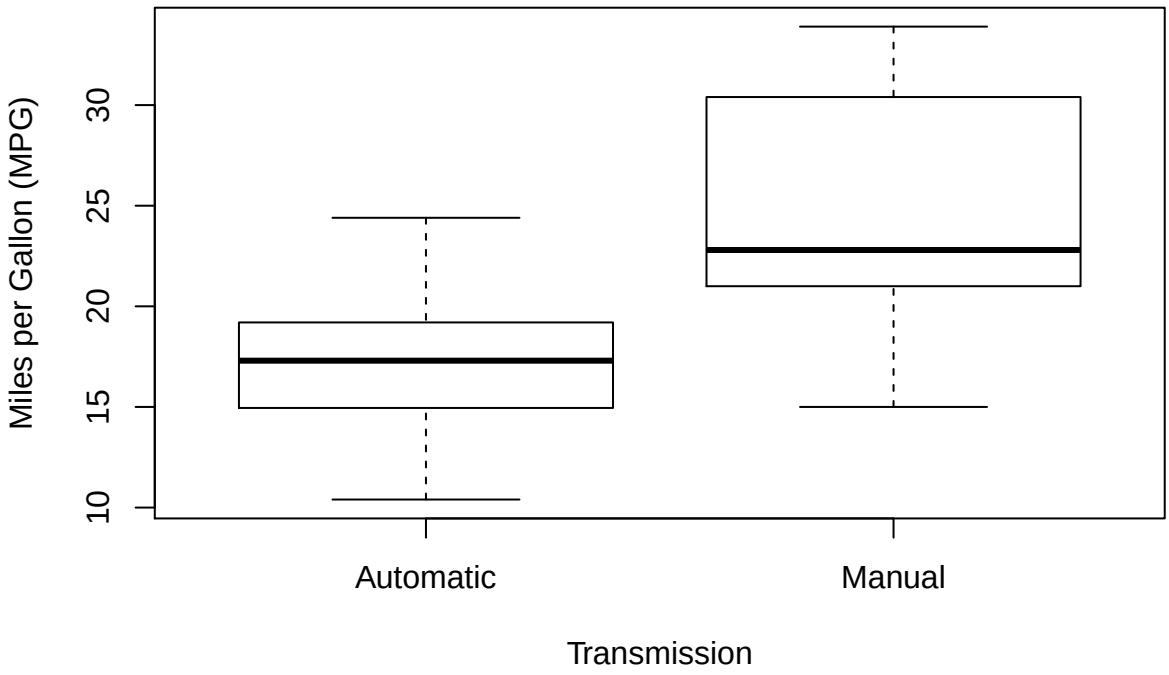

Table 1.2

Correlation Matrix

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.000	-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.600	0.480	-0.551
cyl	-0.852	1.000	0.902	0.832	-0.700	0.782	-0.591	-0.811	-0.523	-0.493	0.527
disp	-0.848	0.902	1.000	0.791	-0.710	0.888	-0.434	-0.710	-0.591	-0.556	0.395
hp	-0.776	0.832	0.791	1.000	-0.449	0.659	-0.708	-0.723	-0.243	-0.126	0.750
drat	0.681	-0.700	-0.710	-0.449	1.000	-0.712	0.091	0.440	0.713	0.700	-0.091
wt	-0.868	0.782	0.888	0.659	-0.712	1.000	-0.175	-0.555	-0.692	-0.583	0.428
qsec	0.419	-0.591	-0.434	-0.708	0.091	-0.175	1.000	0.745	-0.230	-0.213	-0.656
vs	0.664	-0.811	-0.710	-0.723	0.440	-0.555	0.745	1.000	0.168	0.206	-0.570
am	0.600	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168	1.000	0.794	0.058
gear	0.480	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	1.000	0.274
carb	-0.551	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274	1.000

Table 2.1

Model 1 Summary

```
Call:
lm(formula = mpg ~ am, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.392 -3.092 -0.297  3.244  9.508

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    17.15       1.12   15.25  1.1e-15 ***
amManual        7.24       1.76    4.11  0.00029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.9 on 30 degrees of freedom
Multiple R-squared:  0.36,  Adjusted R-squared:  0.338
F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Table 2.2

Model 2 Summary

```
Analysis of Variance Table

Model 1: mpg ~ am
Model 2: mpg ~ am + wt
Model 3: mpg ~ am + wt + cyl
  Res.Df RSS Df Sum of Sq    F  Pr(>F)
```

```
1     30 721
2     29 278  1       443 64.9 9.1e-09 ***
3     28 191  1        87 12.8  0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2.3

Model 2 Summary

```
Call:
lm(formula = mpg ~ am + cyl + disp + hp + drat + wt + vs, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.660 -1.678 -0.417  1.371  5.313

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.4567     9.0238    3.60   0.0014 **
amManual      1.9520     1.7567    1.11   0.2775
cyl          -0.6399     0.8967   -0.71   0.4824
disp          0.0135     0.0121    1.11   0.2769
hp           -0.0303     0.0147   -2.06   0.0500 .
drat          0.5470     1.5101    0.36   0.7204
wt           -3.2453     1.1675   -2.78   0.0104 *
vs            1.3976     1.8484    0.76   0.4569
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.57 on 24 degrees of freedom
Multiple R-squared:  0.859, Adjusted R-squared:  0.818
F-statistic: 20.9 on 7 and 24 DF,  p-value: 9.09e-09
```

Table 3.1

Final Model Summary

```
Call:
lm(formula = mpg ~ wt + hp + I(hp^2) + I(wt^2), data = data)

Residuals:
   Min     1Q Median     3Q    Max
-2.885 -1.817 -0.392  1.350  4.581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.95e+01   3.52e+00   14.04  6.3e-14 ***
wt          -9.22e+00   2.27e+00   -4.06  0.00038 ***
```

```
hp          -9.43e-02   3.19e-02   -2.95  0.00646 **
I(hp^2)      1.74e-04   8.07e-05    2.16  0.03988 *
I(wt^2)      8.50e-01   3.00e-01    2.83  0.00870 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.13 on 27 degrees of freedom
Multiple R-squared:  0.891, Adjusted R-squared:  0.875
F-statistic:   55 on 4 and 27 DF,  p-value: 1.36e-12
```

Figure 3.1: Diagnostic Plot Model 5

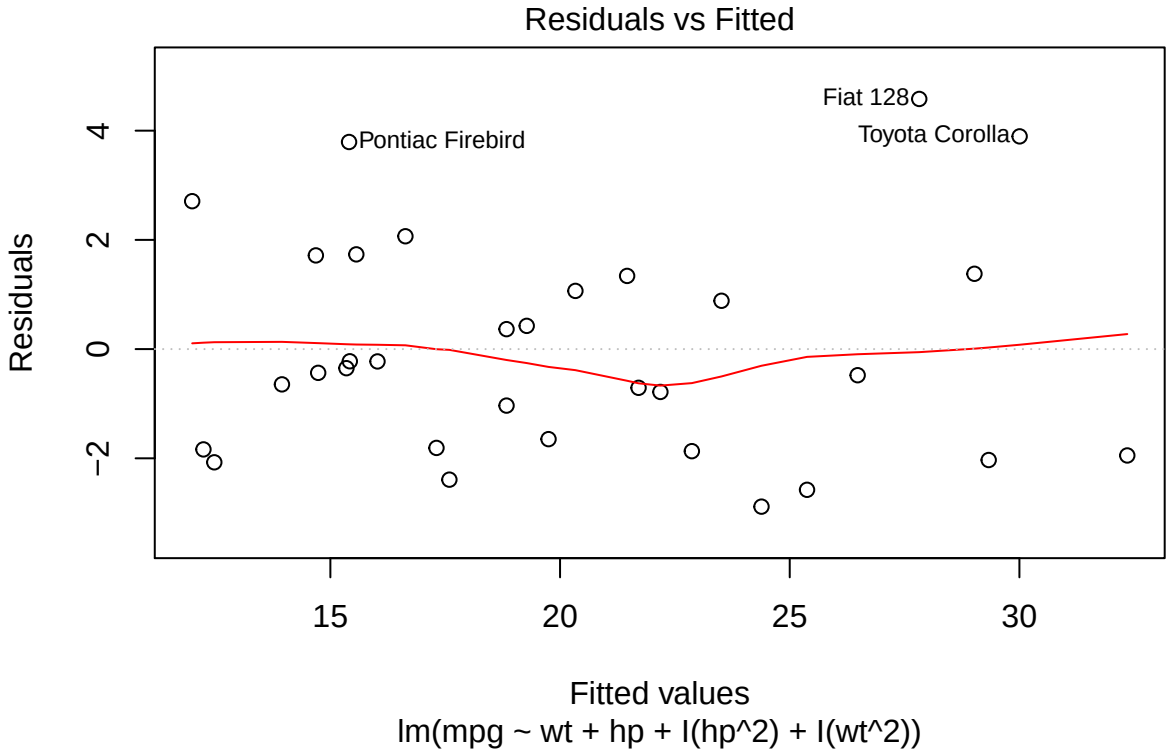

Figure 3.2